Title: The Necessity of Realistic Synthetic Health Data Development Environments

Solution name: Particle Health Sandbox Environment

Organization: Particle Health

Submitter(s): *name*: Parker Bannister
email: parker.bannister@particlehealth.com
mobile: (206) 240-3307
Address: 335 Madison Ave, New York City, New York, 10017

Challenge Category: Category II Entry - Novel Uses of Synthea Generated Synthetic Data

GitHub Repository: https://github.com/ParticleHealth/Particle-Health-Sandbox-Environment

YouTube Video Demonstration: https://youtu.be/Yv3yAXxyiw4

The Necessity of Realistic Synthetic Health Data Development Environments

Abstract

With the ONC's release of the 21st Century Cures Act Final Rule and TEFCA, the healthcare industry will transition to APIs and federal network infrastructures as a necessary means for healthcare data exchange. The intention of this policy is to improve interoperability and allow for new innovation to be built on top of these APIs and federal networks. However, proper tools cannot be built without quality data for testing. Currently, no testing environments populated with synthetic health data that closely model real patient data exist, though they are imperative to enabling the next generation of innovation with a risk-free, low-cost development experience. As a solution, Particle Health has developed the Particle Health Sandbox: a more realistic testing environment, accessed via API, that leverages Synthea's synthetic data to model data from federal health networks that TEFCA refers to. With experience creating a production environment using real patient data from these national health networks, Particle Health is well equipped to address the flaws of current synthetic testing environments through our solution. There are no practical synthetic health data testing environments with CCDA documents and in-document provider notes. Our synthetic health data environment includes modified Synthea generated health records in the form of CCDA documents with accompanying point-in-time documents, in-document chart notes, FHIR documents, as well as an assortment of pre-loaded patient types from Synthea modules, including opioid and COVID-19 patients. Validation through methodical comparison to our production environment and use of CCDA scoring tools ensures that the data is realistic in its variability and quality. Within the risk-free simulated environment of Particle Health's sandbox, developers are enabled to tackle complex care issues, from the opioid epidemic to the COVID-19 pandemic, by leveraging Synthea data to test research and applications with the realism in quality and variability that they can expect from API access or federal network style data, all before *seamlessly* transitioning to real patient data. This level of realism will effectively prepare developers to bring their solutions to fruition faster and with greater accuracy.

Introduction

As a part of the 21st Century Cures Act, The Office of the National Coordinator for Health Information Technology (ONC) delivered a final rule mandating that HL7 FHIR Application Programming Interface (API) capabilities for all health IT developers must comply with the ONC's condition of certification requirements by April 5th of 2021[1]. Additionally, the ONC released Draft 2 of The Trusted Exchange Framework and Common Agreement (TEFCA), which specifies a common set of principles to enable nationwide exchange of electronic health information across disparate health information networks[2]. TEFCA includes APIs as a modality within their framework as a means to facilitate the transmission of healthcare information[2]. Between the ONC's legislation and outlined principles, the healthcare industry is experiencing a push for APIs to be the tool that drives national access to clinical data in a secure manner.

With policy promoting the development and adoption of APIs and national health information networks being the future of healthcare information exchange, application developers and researchers will significantly benefit from improved interoperability. Easy integration with sources of health information will enable increased efficiency across established use cases, like value-based care, and a new wave of applications and research that was previously not possible. National data exchanges are already transmitting billions of patient documents across networks like Carequality, CommonWell and the

eHealth Exchange, but testing environments without protected health information (PHI) simply don't exist. Despite the clear potential, this policy's goal of enabling innovation is hindered by the lack of PHI free settings. Without PHI free environments to access via API, developers and researchers are forced to use live data, making the process of building and testing more precarious. In order for the policy to maximize its intended benefit, realistic PHI free environments must be created for developers to use without high cost or concern over violating HIPAA patient privacy laws. Such developer environments will enable cost-effective, rapid development of applications and research to combat public health issues or Patient Centered Outcome Research.

Synthea is an open source synthetic health data generator developed by the MITRE Corporation[3]. It serves as an excellent tool to generate synthetic health information for the purpose of populating an environment with realistic, yet non-real patient data. Currently, MITRE has a FHIR testing environment that is populated with Synthea generated patient records. However, when queried with an API, their environment responds with FHIR data that does not contain human error or provider written chart notes. This quality of data in the testing environment makes it difficult, if not impossible, to build real world solutions[4]. Real patient data that is currently returned from national health information networks that TEFCA refers to, does however, include human error and variability in entry, provider written chart notes, and is still being returned as CCDA documents with various accompanying point-in-time clinical documents for the time being.

Objective

With the notable absence of risk-free testing environments and the vast difference between the current synthetic patient records returned from testing environments and real patient records returned from the national networks, Particle Health's objective was to construct the Particle Health Sandbox: a more realistic synthetic health data testing environment. With our experience constructing an API that returns real patient information from such national networks, as well as working closely with health technology vendors, we were poised to develop a synthetic environment that can be interacted with through API protocol to closely resemble a realistic health network experience.

Our sandbox leverages Synthea in a novel way with the ability to return FHIR documents, CCDA documents with accompanying point-in-time CCDAs, as well as synthetically generated provider chart notes within these documents. Currently, there are no environments that can be accessed via API that return documents with these added realistic features. Additionally, we have preloaded patients of certain demographics, including patients with opioid usage data, patients with COVID-19 data, as well as other disease types from pre-existing Synthea modules[5, 6, 7].

The realism of our sandbox has been validated in a few ways. We measured the variability of files returned from our production environment, and files returned from national health information networks. Then, we compare this variability to the point-in-time documents we've generated for our sandbox using Synthea. We have also validated these files by leveraging online tools provided by the government to evaluate the document's syntax, quality, and precision compared to those used in the real world[8].

State of the Art Significance

Having a sandbox with data that closely resembles what is returned from national health information networks allows researchers and developers the opportunity to start building without BAAs, risk, or long lead times to combat pressing public health issues. Our sandbox allows developers to

seamlessly transition their development environment from synthetic data to real patient data with PHI once ready. This use of Synthea generated synthetic health information cultivates an environment for the iterative improvement and widespread use of synthetic health information for the benefit of the health and wellbeing of individuals and communities on a national scale.

Realistic features and preloaded patient types further enable testing of innovative solutions. The ability to generate populations with conditions of interest, allows for development to begin right away, rather than parsing through hundreds, if not thousands, of CCDA documents to find synthetic patients with the desired conditions. If it is FHIR that developers prefer, our API supports functionality to search for patients in our sandbox with COVID-19, opioid usage, or other conditions. Synthetic chart notes within the documents allow for a more realistic experience; real patient data often includes many notes that are valuable to the development of Natural Language Processing applications or research.

Particle's synthetic health data sandbox enables companies creating applications for opioid usage or complex care situations, such as COVID-19, to begin development with realistic Synthea generated data. With a pre-loaded opioid patient population, developers or researchers aiming to help manage these patients' conditions can avoid risk or high cost with synthetic data that closely resembles what is returned from national networks. Opioid researchers can leverage point-in-time documents to find lab result documents and gain context about the patient's experience across their care continuum. As importantly, these researchers and developers can easily transition to real patient data without the suprise or disturbance of data with human error, provider notes, and an overbearing amount of returned files. With a high quality testing environment, developers are empowered to iterate their innovations affordably, quickly, and without risk.

Methods

The development of Particle Health's Sandbox followed a stringent process in order to build, validate, and meet challenge requirements. The first stage of development was to create a functional sandbox API that has the same protocol and rule-based-standards as our proprietary production API (that returns real patient information from national health information networks).

To replicate our production API, we leveraged the Google Cloud Platform products, App Engine, and Cloud Storage as open source solutions for our API gateway that responds with our Synthea generated data. We mapped the demographics of our synthetic patients to the Storage Buckets holding respective patient data so that when a request is made for a particular patient it returns the right information. As this is a proprietary process within the Google Cloud Platform, the API setup is not included in our final GitHub repository.

The second stage, and main portion of development, was to generate synthetic healthcare information using Synthea. For this we utilized standard Synthea libraries downloaded from their public GitHub Repo[10]. We modified the provided 'synthea.properties' file to produce CCDA in addition to matching FHIR documents, generate provider notes for these synthetic documents, split the records to obtain more CCDA documents per synthetic patient, as well as output csv and prevalence reports to gain insight on the conditions of the patients generated.

To create point-in-time CCDA documents, we wrote a Python script that templates the new documents with the Python libraries Jinja2 and LXML[11, 12]. This script takes portions of the holistic Continuity of Care CCDA documents Synthea outputs and templates an encounter from their output into a new point-in-time 'Encounter_Summary' CCDA, takes the immunization data and templates that into an 'Immunizations_Summary' CCDA, and follows the same templating pattern for Lab and Medication

Refill Summary documents as well. The clinical notes associated with the patient have also been templated into each document according to the proper CCDA syntax. This pattern yields many point-in-time files for each CCDA Synthea generates with in-document synthetic provider notes.

To produce populations of synthetic patients for our environment with specified diseases or conditions, we have leveraged Synthea's pre-existing disease modules. For opioid and COVID-19 patients, we used the modules for opioid addiction and covid19, respectively, without the addition of any other modules. As not all patients produced had the disease of interest, we generated a large population and used the exported symptoms CSV and prevalence reports, that we previously specified in 'synthea.properties' to output, in order to obtain only the patients with the conditions of interest. We used regular expression to find the patient IDs of patients with Opioid or COVID-19 data from the CSV symptom report. We then used these patient IDs to save the CCDA and FHIR files for those patients with the symptoms of interest. With the patient files of interest, we ran their CCDAs through our templating algorithm and uploaded the produced point-in-time documents with notes to Google Storage Buckets, and loaded their FHIR documents directly into our FHIR store[9].

To validate our environment's synthetic data, we compared the prevalence of the document types we generated to the prevalence of document types that are returned from our production environment. Because there are hundreds of variations and types of point in time documents returned from the national networks, we decided to template the most common documents that accounted for about 66% of all types returned. In addition, we validated the point-in-time documents we generated with our templating algorithm using HealthIT.gov's open source CCDA 2.0 scorecard API, which comprehensively scored our generated files for the quality and precision compared to those used in the real world[8]. We used the CCDA scorecard API to score each point-in-time file we generated and output the results to a CSV to provide a comprehensive view of the results for the entire batch generated. The average score of documents returned from the national network was a C on a scale of A-F, so we wanted to have our output point-in-time documents match this grade for realistic quality.

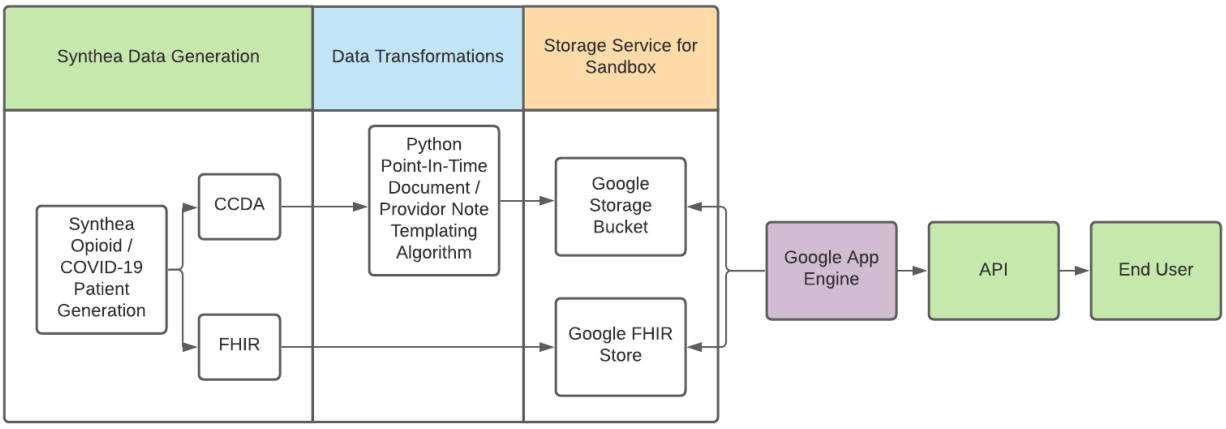

Figure 1 *- A swimlane diagram depicting the stages of development described in the methods section. The first lane shows the use of Synthea for patient generation. The second lane shows the use of Python scripting to create the point-in-time documents and addition of synthetic chart notes into the CCDA documents. The third lane shows the sandbox data being stored in GCP for access. The outside box depicts Google App Engine hosting our API service and environment interaction by users.*

Results

The Particle Health Sandbox we have developed is currently the only realistic testing environment containing point in time CCDA files, FHIR, and in document chart notes that can be accessed via API. Leveraging Synthea's synthetic data we were able to closely model patient data from federal health networks that TEFCA outlines. Our solution bolsters Synthea's data by adding the most common point-in-time documents found in federal networks, such as Encounters, Medication Refills, Immunizations and Lab Summaries, as well as in-document provider notes. This represents a significant improvement to the current Synthea output, which is a single Continuity of Care Document or equivalent in FHIR with separate files for free text notes. This previously constrained developers transitioning from a testing environment to national network data, due to the lack of context and realistic complexities described.

As part of the data generation process, our solution automates an implementation of HealthIT.gov's CCDA scorecard validator. Again, this represents a significant improvement to previous Synthea data, which has not previously been able to validate single CCDA documents. The resulting output of our validation process is a .csv file for each document with the validation results. Notably, we have chosen to benchmark the validity of our data returned to the average score of real patient data, a C rating. While the validity of this data could be improved, in practice, having data in a testing environment that resembles national average scores will assist developers downstream when they eventually move into production. Currently, Synthea data is relatively precise and clean, returning CCDAs that conform to a particular template and doesn't reflect the reality of complex and less validated patient data being returned by the network.

Our Sandbox overcomes a critical limitation of Synthea's in its inability to produce patients with a specified condition. While this may be the intention of Synthea data, to reproduce a realistic random sample, our solution provides developers with the ability to focus on a subset of patients. Our sandbox achieves this by using regular expression and supporting symptom .csv's that Synthea produces to find and retrieve patients with IDs relating to particular conditions or symptoms of interest.

Our solution was designed in the spirit of reusability and reproducibility and we hope that other developers will access and utilize it to its fullest capabilities. We have used open source tools and libraries for the solution, including the validation process, and it is published, in its entirety, on GitHub. A Readme is attached for instructions on how to use the solution, as well as the requirements needed to run.

Ultimately, this solution aims to improve the clinical relevance of Synthea data so that it might enhance the development of future innovations in healthcare. In generating realistic, point-in-time CCDA data, across a patient's continuum of care, our solution allows developers to plan for realistic data extraction and utilization and ultimately to more effectively improve healthcare outcomes, both clinical and otherwise. The point in time documents we have generated reflect the most common and available documents returned, based on Particle Health's extensive experience in calling data from the federal networks through our production API.

References

1) “Information Blocking and the ONC Health IT Certification Program: Extension of Compliance Dates and Timeframes in Response to the COVID-19 Public Health Emergency.” Federal Register, 4 Nov. 2020, www.federalregister.gov/documents/2020/11/04/2020-24376/information-blocking-and-the-onc-health-it-certification-program-extension-of-compliance-dates-and.
2) “Trusted Exchange Framework and Common Agreement.” HealthIT.gov, 26 May 2020, www.healthit.gov/topic/interoperability/trusted-exchange-framework-and-common-agreement.
3) Synthea.mitre.org, synthea.mitre.org/about.
4) Synthea.mitre.org, synthea.mitre.org/fhir-api.
5) Syntheticheath. “Opioid Addiction Module · Issue #511 · Syntheticheath/Synthea.” GitHub, github.com/syntheticheath/synthea/issues/511.
6) Walonoski J, Klaus S, Granger E, et al. Synthea™ Novel coronavirus (COVID-19) model and synthetic data set. Intell Based Med. 2020;1:100007. doi:10.1016/j.ibmed.2020.100007
7) Synthea Generic Module Builder, syntheticheath.github.io/module-builder/.
8) “C-CDA Scorecard.” SITE, site.healthit.gov/scorecard/.
9) “FHIR | Cloud Healthcare API | Google Cloud.” Google, Google, cloud.google.com/healthcare/docs/concepts/fhir.
10) Syntheticheath. “Syntheticheath/Synthea.” GitHub, github.com/syntheticheath/synthea.
11) “Jinja.” Jinja, jinja.palletsprojects.com/en/2.11.x/index.html.
12) “XML and HTML with Python.” Lxml, lxml.de/.